2010

Is transcription factor binding site turnover a sufficient explanation for cis-regulatory sequence divergence?

Sandeep Venkataram
Washington University School of Medicine in St. Louis

Justin C. Fay
Washington University School of Medicine in St. Louis

GBE

Genome Biology and Evolution

SMBE

Is Transcription Factor Binding Site Turnover a Sufficient Explanation for Cis-Regulatory Sequence Divergence?

Sandeep Venkataram[1,2], and Justin C. Fay*[,1,2]

[1]Department of Genetics, Washington University, St. Louis, Missouri

[2]Center for Genome Sciences, Washington University, St. Louis, Missouri

*Corresponding author: E-mail: jfay@genetics.wustl.edu.

Accepted: 10 October 2010

Abstract

The molecular evolution of cis-regulatory sequences is not well understood. Comparisons of closely related species show that cis-regulatory sequences contain a large number of sites constrained by purifying selection. In contrast, there are a number of examples from distantly related species where cis-regulatory sequences retain little to no sequence similarity but drive similar patterns of gene expression. Binding site turnover, whereby the gain of a redundant binding site enables loss of a previously functional site, is one model by which cis-regulatory sequences can diverge without a concurrent change in function. To determine whether cis-regulatory sequence divergence is consistent with binding site turnover, we examined binding site evolution within orthologous intergenic sequences from 14 yeast species defined by their syntenic relationships with adjacent coding sequences. Both local and global alignments show that nearly all distantly related orthologous cis-regulatory sequences have no significant level of sequence similarity but are enriched for experimentally identified binding sites. Yet, a significant proportion of experimentally identified binding sites that are conserved in closely related species are absent in distantly related species and so cannot be explained by binding site turnover. Depletion of binding sites depends on the transcription factor but is detectable for a quarter of all transcription factors examined. Our results imply that binding site turnover is not a sufficient explanation for cis-regulatory sequence evolution.

Key words: evolution, regulation, yeast.

Introduction

Most of our understanding of molecular evolution comes from the analysis of protein coding sequences (Li 2006), which are often highly conserved in both sequence and function between closely and even distantly related species (Tatusov et al. 2003). In contrast, cis-regulatory sequences are much more labile. Although comparison of closely related species shows that there are just as many conserved noncoding as coding sequences within a genome (Siepel et al. 2005), comparison of distantly related species shows that only a small fraction of noncoding sequences conserved in closely related species are also conserved in distantly related species, for example (Margulies et al. 2005; Woolfe et al. 2005). In a number of cases, gene regulation is conserved despite the absence of conservation at the primary sequence level (Tautz 2000; Weirauch and Hughes 2010).

Binding site turnover provides one explanation for divergence in sequence without a concomitant change in gene regulation (Hancock et al. 1999; Ludwig et al. 2000; Dermitzakis and Clark 2002). In this scenario, the gain of a functionally redundant transcription factor binding site enables a previously conserved binding site for the same transcription factor to be lost. Comparative genomic analysis of experimentally identified binding sites provides substantial evidence for binding site turnover in a number of different species (Dermitzakis and Clark 2002; Costas et al. 2003; Dermitzakis et al. 2003; Moses et al. 2006; Doniger and Fay 2007; Otto et al. 2009; Bradley et al. 2010).

Divergence in transcriptional regulation can also result in the absence of conserved cis-regulatory sequences. There is a growing number of examples in which orthologous transcription factors have been shown to regulate different sets of genes (Tsong et al. 2003; Ihmels et al. 2005; Tanay et al. 2005; Tsong et al. 2006; Borneman et al. 2007; Martchenko et al. 2007; Odom et al. 2007; Hogues et al. 2008; Tuch, Galgoczy, et al. 2008; Perez and Groisman 2009b; Schmidt et al.

2010). These studies support a model of transcriptional rewiring whereby homologous genes are regulated by different transcription factors (Tuch, Li, and Johnson 2008; Lavoie et al. 2009; Perez and Groisman 2009a). Although gene regulation can be conserved through substitution of one transcriptional regulator for another, transcriptional rewiring may also involve divergent regulatory outputs (Ihmels et al. 2005; Brown et al. 2009; Lavoie et al. 2009; Perez and Groisman 2009a). The transcription rewiring model is distinct from that of binding site turnover because the later does not involve changes in the set of genes regulated by a transcription factor.

The extent to which binding site turnover can explain the lack of sequence similarity between distantly related species has been difficult to assess. First, orthologous cis-regulatory sequences are not easy to identify unless they show some level of sequence similarity. Second, transcription factors bind short sequences that are often present once every thousand bases in the genome. Thus, even when two orthologous cis-regulatory sequences have been identified, it is difficult to know whether the presence of a binding site in both sequences is due to binding site turnover or chance.

To determine whether binding site turnover is consistent with cis-regulatory sequence divergence, we compared the presence and absence of binding sites across a diverse set of 14 yeast genomes. Yeast have short, typically 500 bp, intergenic sequences that facilitate the identification and analysis of binding site evolution. We generated a set of orthologous intergenic sequences irrespective of the sequence similarity based on their syntenic relationships with adjacent coding sequences. By examining the conservation of binding sites identified in *Saccharomyces cerevisiae*, we found that while some transcription factor's binding sites are consistent with a model of binding site turnover, a quarter of the transcription factors are consistent with some amount of regulatory divergence.

Materials and Methods

Identification of Syntenic Intergenic Regions

Sequences for the 14 species used in this study (*S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *S. castelli*, *Candida glabrata*, *Kluyveromyces polysporus*, *Zygosaccharomyces rouxii*, *K. thermotolerans*, *K. waltii*, *S. kluvyerii*, *K. lactis*, *Ashbya gossypii*) were obtained from the Saccharomyces Genome Database (SGD) and the Ashbya Genome Database on 8 November 2007 and from the Wolfe lab's genome browser on 7 March 2009. The *S. cerevisiae* gene annotations (SGD_features.tab) was obtained from SGD on 8 November 2007. Every open reading frame defined in the annotation file was found in the *S. cerevisiae* genome and used to identify homologous protein coding sequences using TBlastX (WU-BLAST 2.0MP) with an *E*-value cutoff set to 10^{-10}, a query frame set to 1, an hspsepSmax set to 10,000 and a seg filter. Intergenic regions syntenic to an *S. cerevisiae* intergenic region were defined flanking homologous genes in the same relative orientation as in *S. cerevisiae* and having an intergenic region within 3-fold of the size of the corresponding intergenic region in *S. cerevisiae*. In the case of multiple possible syntenic regions between *S. cerevisiae* and a given species, we chose the one with the lowest summed Blast *E*-value. The intergenic regions in species other than *S. cerevisiae* were defined based on *S. cerevisiae* gene annotations and global alignments of both intergenic and flanking coding sequences.

Global Alignment of Syntenic Intergenic Regions

The Needleman–Wunsch algorithm was used to generate pairwise alignments between each *S. cerevisiae* intergenic region with the syntenic region found in each of the other species. Flanking protein coding sequences were included in the alignments, and percent identity was calculated using intergenic regions defined in *S. cerevisiae*. A gap open penalty of 6 and a gap extension penalty of 0.2 were used. MCALIGN2 (Wang et al. 2006) was also used to generate pairwise alignments. A custom insertion/deletion rate was used based on data from three closely related *S. cerevisiae* strains (Doniger et al. 2008). The relative rate of point substitutions to insertion/deletions was set to 6 and the relative frequency of 1, 2, 3, etc bp insertion/deletions was set to 0.62, 0.18, 0.06, 0.05, 0.02, 0.03, 0.01, 0.01, 0.01, 0.01. Alignments are available upon request from the corresponding author.

Significant Similarity between Syntenic Intergenic Regions

BlastN (WU-BLAST 2.0MP) and HMMER (v2.0) were used to search each genome for similarity to *S. cerevisiae* intergenic regions. For this analysis, only intergenic regions were used that were upstream of a gene, that is, convergently transcribed intergenic regions were removed. For BlastN, significant similarity was defined by an *E*-value cutoff of 10^{-10}, hspsep max = 10,000, and for HMMER, significant similarity was defined by an *E*-value cutoff of 10^{-10}. HMMER is a profile alignment algorithm and was trained on *sensu strictu* species intergenic sequences (*S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*) aligned using ClustalW and then run on each genome not included in the training alignments.

Identification of Transcription Factor Binding Sites

Experimentally identified transcription factor binding sites were obtained for 2,622 syntenic intergenic regions based on chromatin immunoprecipitation experiments involving 126 transcription factors (Harbison et al. 2004). Only syntenic intergenic regions containing promoters were used. Using a *P* value cutoff of 0.005 for significant binding, we used a total of 6,459 binding sites for 118 transcription factors which bound at least one of the *S. cerevisiae* syntenic intergenic regions. For each bound intergenic region, the orthologous intergenic regions were searched for binding sites using Patser and position weight matrices derived from

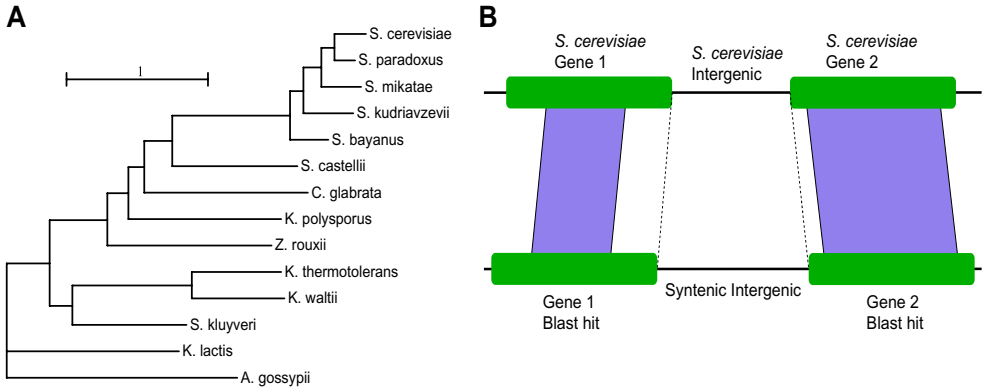

Fig. 1.—Identification of syntenic intergenic regions. (*A*) A maximum likelihood tree of 14 yeast species used to identify syntenic intergenic sequences. The tree is based on concatenation of 13 genes (YMR009W, YLR147C, YJR034W, YLR029C, YIL074C, YHR142W, YGR284C, YCL055W, YJL072C, YCR036W, YOR250C, YBR196C, YBR282W) for which homologs were identified in all species. Branch lengths show the synonymous substitution rate calculated using HYPHY and model MG94xHKY85. (*B*) Syntenic intergenic regions were defined by homology of adjacent protein coding sequences (blue). Intergenic sequences were defined using the ends of the protein coding sequences as annotated in *S. cerevisiae*.

the binding data (MacIsaac et al. 2006). Our initial analysis showed that no significant matches were found in many *S. cerevisiae* bound regions due to the stringency of the default Patser cutoff. To avoid missing binding sites due to overly stringent cutoffs, we used a minimum ln(*P* value) cutoff of −10 calculated from the log likelihood of the motif versus background sequence using the information content of the motif (Hertz and Stormo 1999). Running this on *S. cerevisiae* intergenics, we identified binding sites for 60% of the regions found to be bound by a particular transcription factor. Binding sites were also identified using the same method for orthologous intergenic regions for a set of 15 promoter regions that were carefully characterized by promoter bashing, footprinting, EMSA, or mutation analysis (supplementary table 1, Supplementary Material online).

Simulated and Randomized Intergenic Sequences

Intergenic sequences were randomized by selecting sites without replacement. Simulations of intergenic sequences were performed using the CisEvolver software package that evolves a sequence according to a specified tree and substitution rate and returns the resulting evolved sequences (Pollard et al. 2006). The tree and synonymous substitution rate were obtained from 13 genes with data from all species (fig. 1). The tree was re-rooted, such that *S. cerevisiae* was at the root and we used the *S. cerevisiae* intergenic as the starting input sequence. Insertion/deletion rates and length distributions were the same as those used for MCALIGN. A total of 10 randomized and 10 simulated sequences were generated for each intergenic region.

Results

To identify orthologous intergenic sequences from 14 yeast species, we searched for sequences with homology to adjacent protein coding sequences in *S. cerevisiae*. Syntenic intergenic regions were defined by two open reading frames in the same relative orientation in both species and within 3-fold of the *S. cerevisiae* intergenic size (fig. 1). Using TBlastX to establish homology between open reading frames, we identified 28,182 regions from 13 species syntenic to one of 5,957 intergenic regions in *S. cerevisiae*. The number of syntenic intergenic regions declined with increasing distance from *S. cerevisiae* but remained relatively constant outside of the more closely related *sensu strictu* Saccharomyces species (table 1). Relative to *S. cerevisiae*, the median GC content and intergenic length were similar in most species. However, *K. thermotolerans*, *K. waltii*, and *A. gossypii* showed a GC content 5% higher than *S. cerevisiae* and *K. thermotolerans*, *K. lactis* and *Z. rouxii* showed a median intergenic length greater than four times that of *S. cerevisiae* (table 1).

To compare sequence similarity among syntenic intergenic regions, we used 1,065 regions with syntenic homologs in nine or more species. Using the Needleman–Wunsch algorithm, we aligned the entire syntenic region, including both flanking coding regions between *S. cerevisiae* and each of the other species. Figure 2 shows the average percent identity of the 1,065 intergenic regions compared with the percent identify from alignment of randomized intergenic regions. With the exception of the *sensu strictu* Saccharomyces species, the average percent identity was close to 40% and not significantly different from that of randomized intergenic regions. We also calculated percent identity from MCALIGN2 alignments using insertion, deletion, and substitution parameters derived from closely related strains of *S. cerevisiae* (see Materials and Methods). MCALIGN2 alignments showed lower percent identities for each species compared with the Needleman–Wunsch alignments but also showed no significant similarity outside of the *sensu strictu* Saccharomyces species.

Table 1
Characteristics of Syntenic Intergenic Regions

Species	Syntenic Intergenic Regions	Median GC Content	Median Length
S. cerevisiae	5,957	0.35	373
S. paradoxus	4,572	0.35	353
S. mikatae	4,305	0.34	344
S. kudriavzevii	3,822	0.36	353
S. bayanus	4,282	0.37	352
S. castellii	1,808	0.34	298
C. glabrata	1,821	0.35	454
K. polysporus	963	0.30	500
Z. rouxii	1,361	0.37	1,588
K. thermotolerans	1,154	0.46	1,592
K. waltii	1,094	0.42	350
S. kluyveri	939	0.39	374
K. lactis	1,039	0.36	3,176
A. gossypii	1,022	0.51	373

Although the majority of intergenic regions showed no significant sequence similarity between distantly related species, there may be a small subset of syntenic orthologs that have high levels of sequence similarity across a portion of the intergenic region. To identify significant sequence similarity between distantly related intergenic regions, we used the local alignment algorithm, BlastN, and a profile hidden markov alignment algorithm, HMMER, to search the genome of each species for similarity to each *S. cerevisiae* intergenic sequence. With the exception of the *sensu strictu* Saccharomyces species, BlastN identified fewer than 2% of syntenic intergenic regions as showing significant similarity (fig. 3). Those intergenic regions identified by BlastN typically contained small regions of high sequence similarity and an average percent identity over the entire intergenic region of greater than 60% (supplementary fig. 1, Supplementary Material online). When trained on alignments of the *sensu strictu* Saccharomyces species, HMMER identified a small but slightly higher percentage of syntenic intergenic regions (fig. 3). Thus, little sequence similarity remains between distantly related orthologous intergenic regions.

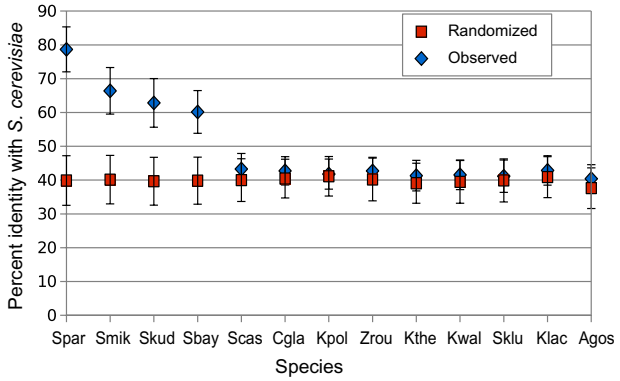

Fig. 2.—Intergenic regions from distantly related species show an average percent identity that is not significantly greater than that of randomized intergenic regions. The percent identity including gaps from Needleman–Wunsch alignments of each species with *S. cerevisiae* (blue) relative to that from alignment of randomized regions with *S. cerevisiae* (red). Average percent identify and standard errors (bars) were calculated from 1,065 intergenic regions with syntenic orthologs in nine or more species.

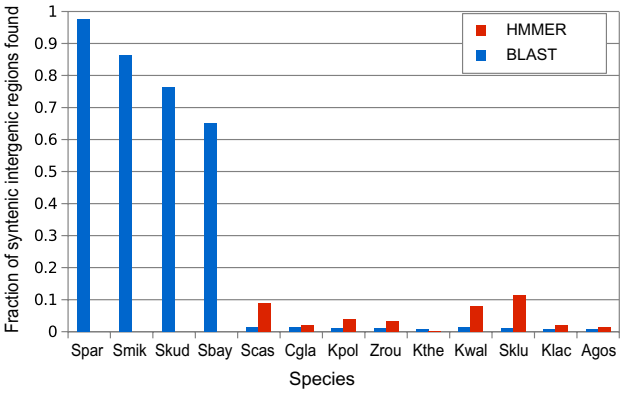

Fig. 3.—Few intergenic regions from distantly related species show significant similarity to syntenic *S. cerevisiae* intergenic regions. The fraction of syntenic intergenic regions found by BlastN searches (blue) and HMMER searches (red) of each species' genome using *S. cerevisiae* intergenic sequences as a query.

Turnover of transcription factor binding sites provides a simple model whereby the function of distantly related promoters can be conserved while their sequences diverge (Hancock et al. 1999; Ludwig et al. 2000; Dermitzakis and Clark 2002). If the lack of sequence similarity between distantly related orthologous promoter regions can be explained by binding site turnover, experimentally identified binding sites in *S. cerevisiae* should also be present within orthologous cis-regulatory sequences, although not necessarily in the same position or orientation.

To determine how often transcription factor binding sites in *S. cerevisiae* are also present in distantly related orthologous intergenic sequences, we used a set of 6,459 binding sites identified for 118 transcription factors in *S. cerevisiae* based on chromatin immunoprecipitation experiments (Harbison et al. 2004; MacIsaac et al. 2006). For each binding site, a position weight matrix model of the binding site was used to search each orthologous intergenic sequence. Figure 4 shows that there is a significant enrichment of binding sites in orthologous intergenic sequences compared with randomized and simulated intergenic sequences for each species. We used simulated intergenic sequences based on synonymous site divergence within coding sequences to control for the lack of divergence expected over short evolutionary time periods. The frequency of binding sites in the simulated sequences is close to that of the randomized sequences for all species except *S. paradoxus* (19% vs. 12%, respectively), consistent with the high but not saturated synonymous substitution rate of 0.35 substitutions

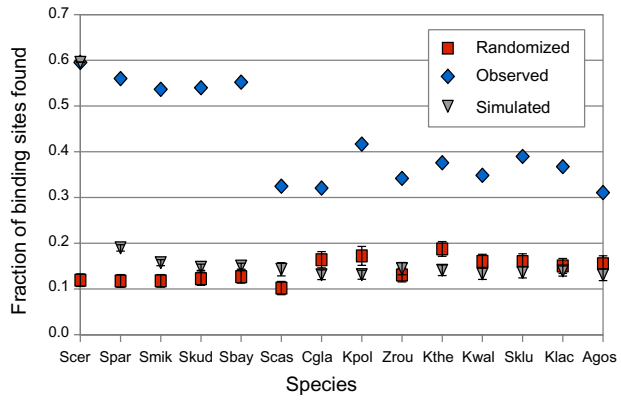

Fig. 4.—Transcription factor binding sites identified in *S. cerevisiae* occur more often than expected by chance in both closely and distantly related syntenic intergenic regions. The fraction of *S. cerevisiae* binding sites found in each species (blue) compared with randomized (red) and simulated (grey) intergenic sequences.

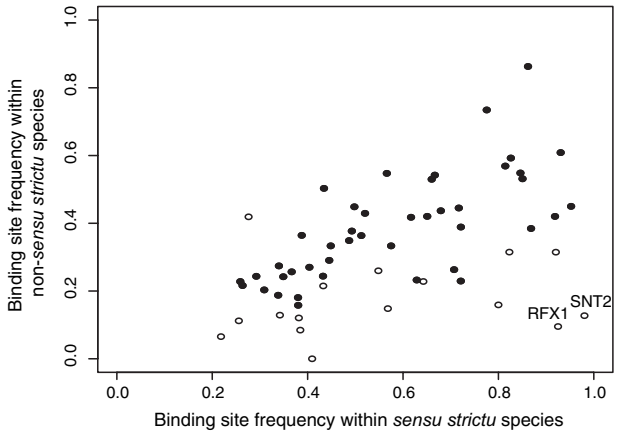

Fig. 5.—Heterogeneity in the frequency of transcription factor binding sites within closely and distantly related species. The frequency of binding sites is shown for 59 transcription factors, open circles show sites that have significant frequency differences between the closely and distantly related species.

per site between *S. cerevisiae* and *S. paradoxus*. The enrichment of binding sites in distantly related species supports the binding site turnover model and implies that at least some binding sites are conserved. However, the distantly related species contained significant fewer binding sites than the *sensu strictu* Saccharomyces species (35% vs. 56%, $P < 0.001$, Mann–Whitney *U* test). Although the percentage of binding sites found in the distantly related species depends on the cutoff used to define a binding site, the distantly related species have fewer binding sites than the *sensu strictu* Saccharomyces species regardless of a more or less stringent cutoff (supplementary fig. 2, Supplementary Material online). This suggests that changes in binding specificity are unlikely to explain the difference between the closely and distantly related species unless binding specificity of the transcription factor is dramatically altered.

The absence of *S. cerevisiae* binding sites in the distantly related species could be the result of a more complex model whereby one binding site is substituted for another site bound by a different transcription factor. However, it is also possible that some of the *S. cerevisiae* binding sites are not functional despite being bound in *S. cerevisiae*. To examine this latter possibility, we used a smaller set of 41 binding sites bound by 18 different transcription factors within 15 promoters. Each of these binding sites has a large effect on gene expression and was identified by promoter bashing, footprinting, gel-shift, or mutation analysis (supplementary table 1, Supplementary Material online). For this small set of carefully annotated binding sites, we found 31% of sites were conserved within the *sensu strictu* Saccharomyces species but a significantly smaller fraction, 26%, were conserved in the distantly related species ($P = 0.019$, Mann–Whitney *U* test). Although the difference between the closely and distantly related species is not as large as that as the larger set of binding sites defined by chromatin immunoprecipitation, the small number of carefully annotated sites combined with their low rates of conservation within the closely related species make it difficult to know whether the two sets of data are different from one another. However, both sets of data suggest that orthologous genes are more often regulated by different transcription factors in the distantly related compared with the closely related species.

Not all binding sites may evolve under the same constraints. Binding sites for some transcription factors may typically evolve through binding site turnover, whereas binding sites for other transcription factors may often be lost, gained, or exchanged for sites bound by another transcription factor. To identity binding sites inconsistent with binding site turnover, we compared the proportion of sites present within the *sensu strictu* Saccharomyces species with the proportion present in the distantly related species for each transcription factor. We excluded *S. cerevisiae* from the *sensu strictu* species and subtracted the number of sites expected by chance based on simulated intergenic regions from the observed number of sites. To avoid small sample sizes, we also excluded 59 of the 118 transcription factors that showed no significant difference between the observed and simulated frequency of binding sites in the *sensu strictu* Saccharomyces species. Of the remaining 59 transcription factors, 43 showed no significant difference in the frequency of binding sites between the closely and distantly related species and 15 (25%) showed a significantly higher proportion of sites in the closely relative to the distantly related species ($P < 0.01$, Fisher's Exact Test, fig. 5). Interestingly, for the 59 Hap2 bound intergenic regions, there were more Hap2 sites found in the distantly related compared with the closely related species. However, with a *P* value cutoff of 0.01, we expect just under one false positive due to testing 59 transcription factors. Transcription factors with

more sites in the close relative to the distant species function in a variety of biological processes, including the cell cycle, pseudohyphal growth, and meiosis. The two transcription factors showing the largest difference in binding site frequency between the closely and distantly related species are Rfx1, involved in response to DNA damage, and Snt2, predicted to play in a role in regulation of amine transporters (Ward and Bussemaker 2008). Thus, although an appreciable number of transcription factors may be rewired to regulate different genes, there is no obvious distinction between these transcription factors and those with predominantly conserved binding sites.

Some binding sites may be involved in regulatory divergence between pre- and postwhole genome duplicated species. In yeast, a whole-genome duplication has been associated with a number of phenotypes related to an increased tendency for aerobic fermentation (Piskur et al. 2006). To compare the frequency of binding sites between the pre- and postwhole genome duplicated species, we excluded the closely related *sensu strictu* Saccharomyces species. Four transcription factors, Abf1, Cbf1, Gln3 and Tye7, show a significant difference in abundance between the pre- and postwhole genome duplicated species ($P < 0.05$, Bonferroni corrected Fisher's Exact Test). Interestingly, only Gln3, involved in nitrogen catabolite repression, has a lower abundance in the postwhole relative to the prewhole genome duplicated species.

Discussion

Divergence in cis-regulatory sequences without a concomitant change in gene regulation presents a significant challenge to understanding gene regulation, evolution of gene regulation and how changes in gene regulation contribute to phenotypic divergence. By identifying orthologous intergenic sequence across a range of yeast species, we show that there is little to no sequence similarity between *S. cerevisiae* and species outside of the *sensu strictu* Saccharomyces clade. Our analysis of binding sites within orthologous cis-regulatory sequences shows that while some transcription factors have binding sites that are equally conserved in both closely and distantly related species, consistent with the binding site turnover model, a quarter of the transcription factors have binding sites that are significant depleted in the distantly related yeast species, consistent with a model of transcriptional rewiring of gene regulation.

Understanding the molecular evolution of cis-regulatory sequences is beset by a number of challenges. First, defining cis-regulatory sequences is not easy. Conservation can be used to identify cis-regulatory sequences but not all cis-regulatory sequences are conserved, for example (Frazer et al. 2004; Prabhakar et al. 2006). This makes it difficult to measure the degree to which cis-regulatory sequences are conserved without circularity. Transcription factor binding can be used to define cis-regulatory sequences but not all binding events are relevant to the organism.

Enhancers that pattern the early Drosophila embryo have been one of the best models for studying the evolution of cis-regulatory sequences because they have well-defined functions under specific conditions (Simpson and Ayyar 2008). However, there is some uncertainty as to whether the results from these early-acting developmental enhancers can be generalized to other cis-regulatory sequences and other species. Our work in yeast complements that done in Drosophila since in yeast cis-regulatory sequences are contained within short intergenic sequences and so do not need to be localized experimentally. By searching orthologous intergenic sequences for a small set of a carefully defined transcription factor binding sites as well as for a larger set of sites defined by chromatin immunoprecipitation experiments in *S. cerevisiae*, we show that a substantial fraction of binding sites are absent in distantly related species and so cannot be explained by binding site turnover. Presumably, many of the cis-regulatory sequences drive similar patterns of gene expression through use of other transcription factors not used by *S. cerevisiae.* However, it is also possible that the absence of these binding sites result in species-specific differences in gene expression.

A second challenge to understanding the molecular evolution of cis-regulatory sequences is that their regulatory output can often be conserved with little or no conservation at the primary sequence level. A number of compelling of examples of such have been shown through use of heterologous expression assays (Tautz 2000; Weirauch and Hughes 2010). However, with only a small number of examples, it is difficult to know whether these observations are particular to certain types of genes and the average time period over which sequence similarity disappears. By using syntenic intergenic regions and global alignments anchored on either side by conserved protein coding sequences, we find that the vast majority of cis-regulatory sequences in *S. cerevisiae* have no significant level of sequence similarity with species outside of the *sensu strictu* Saccharomyces clade. Our results are concordant with another genome study which found conservation of tissue-specific expression is not correlated with conservation of noncoding sequences (Chan et al. 2009) and provide a data set of well-defined orthologous cis-regulatory sequences that can be used to understand gene regulation and its evolution. A key component needed to better interpret these comparisons is a large set of heterologous expression assays from both closely and distantly related species irrespective of sequence conservation.

By comparing binding site conservation of different transcription factors, we find diverse modes of evolution. Some binding sites are as frequent in closely related species as distantly related species, consistent with binding site turnover, whereas others are significantly depleted, consistent with transcriptional rewiring. We found no obvious distinction between these two groups, either in terms of the functions of the transcription factors or information content of the

binding site motifs. Interestingly, all three of the transcription factors with significantly more conserved binding sites within the postwhole genome duplicated compared with prewhole genome duplicated species have been associated with the regulation of glycolytic genes and may be related to the shift in metabolism from respiration to fermentation in the presence of oxygen (Piskur et al. 2006). Both Cbf1 and Tye7 share the same core motif, CACGTG, but bind to different promoters and co-occur with Gcr2 binding sites, known to be involved with the activation of glycolytic genes (Chambers et al. 1995; Gordân et al. 2009). Although Tye7 is specifically involved in the regulation of glycolytic genes (Nishi et al. 1995), Cbf1 binds many loci, including the promoters of methionine metabolism genes and centromeres (Kent et al. 2004). Similarly, Abf1 is involved in DNA replication and repair and regulates genes of diverse function, including glycolytic genes (Chambers et al. 1995). Although Gcr2 binding sites are present at equal frequencies within the prewhole and postwhole genome duplicated species, other well-characterized regulators of glyolytic genes, Mig1, Rgt1 and Gcr1, were not tested due to the small number of bound syntenic intergenic regions.

One drawback of our analysis is that it was not optimized for the identification of binding sites with significant gains or losses along different lineages. First, we limited our analysis to 1,065 syntenic intergenic regions. Second, likelihood-based approaches that test for a constant or accelerated rate of binding site gain/loss would more explicitly test for transcription factors with altered sets of target genes (Otto et al. 2009).

Our results indicate that transcriptional rewiring either with or without divergence in gene expression often contributes to divergence within cis-regulatory sequences. Most evidence for transcriptional rewiring in yeast has been based on two distantly related species, *C. albicans* and *S. cerevisiae* (Tuch, Li, and Johnson 2008; Lavoie et al. 2009). Our results are consistent with the idea that transcriptional rewiring is a general feature of many transcription factors and may often occur over much shorter time periods. Chromatin immunoprecipitation experiments have shown some transcription factors bind largely different sets of genes between closely related species (Borneman et al. 2007; Odom et al. 2007; Bradley et al. 2010; Schmidt et al. 2010) as well as between different individuals of the same species (Kasowski et al. 2010; Zheng et al. 2010). These studies highlight the importance of distinguishing gain or loss of binding sites relevant to species' or individuals' phenotypic differences from those gains and losses that occur by chance.

Supplementary Material

Supplementary figures S1–S2 and table S1 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We thank members of the Fay laboratory and Scott Doniger for feedback and suggestions and two anonymous reviewers for suggesting a number of modifications to our analysis and presentation. This work was supported by the National Institute of General Medical Sciences (grant GM080669 to J.F.) and a Howard Hughes Medical Institute Summer Undergraduate Research Fellowship (to S.V.).

Literature Cited

Borneman AR, et al. 2007. Divergence of transcription factor binding sites across related yeast species. Science. 317:815–819.

Bradley RK, et al. 2010. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. PLoS Biol. 8:e1000343.

Brown V, Sabina J, Johnston M. 2009. Specialized sugar sensing in diverse fungi. Curr Biol. 19:436–441.

Chambers A, Packham EA, Graham IR. 1995. Control of glycolytic gene expression in the budding yeast (*Saccharomyces cerevisiae*). Curr Genet. 29:1–9.

Chan ET, et al. 2009. Conservation of core gene expression in vertebrate tissues. J Biol. 8:33.

Costas J, Casares F, Vieira J. 2003. Turnover of binding sites for transcription factors involved in early Drosophila development. Gene. 310:215–220.

Dermitzakis E, Bergman C, Clark A. 2003. Tracing the evolutionary history of Drosophila regulatory regions with models that identify transcription factor binding sites. Mol Biol Evol. 20: 703–714.

Dermitzakis E, Clark A. 2002. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Mol Biol Evol. 19:1114–1121.

Doniger SW, Fay JC. 2007. Frequent gain and loss of functional transcription factor binding sites. PLoS Comput Biol. 3:e99.

Doniger SW, et al. 2008. A catalog of neutral and deleterious polymorphism in yeast. PLoS Genet. 4:e1000183.

Frazer K, et al. 2004. Noncoding sequences conserved in a limited number of mammals in the SIM2 interval are frequently functional. Genome Res. 14:367–372.

Gordân R, Hartemink AJ, Bulyk ML. 2009. Distinguishing direct versus indirect transcription factor-DNA interactions. Genome Res. 19: 2090–2100.

Hancock JM, Shaw PJ, Bonneton F, Dover GA. 1999. High sequence turnover in the regulatory regions of the developmental gene hunchback in insects. Mol Biol Evol. 16:253–265.

Harbison CT, et al. 2004. Transcriptional regulatory code of a eukaryotic genome. Nature. 431:99–104.

Hertz GZ, Stormo GD. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics. 15:563–577.

Hogues H, et al. 2008. Transcription factor substitution during the evolution of fungal ribosome regulation. Mol Cell. 29:552–562.

Ihmels J, et al. 2005. Rewiring of the yeast transcriptional network through the evolution of motif usage. Science. 309:938–940.

Kasowski M, et al. 2010. Variation in transcription factor binding among humans. Science. 328:232–235.

Kent NA, Eibert SM, Mellor J. 2004. Cbf1p is required for chromatin remodeling at promoter-proximal CACGTG motifs in yeast. J Biol Chem. 279:27116–27123.

Lavoie H, Hogues H, Whiteway M. 2009. Rearrangements of the transcriptional regulatory networks of metabolic pathways in fungi. Curr Opin Microbiol. 12:655–663.
Li W. 2006. Molecular evolution. Sunderland (MA): Sinauer Associates.
Ludwig M, Bergman C, Patel N, Kreitman M. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature. 403:564–567.
MacIsaac KD, et al. 2006. An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. BMC Bioinformatics. 7:113.
Margulies EH, et al. 2005. Comparative sequencing provides insights about the structure and conservation of marsupial and monotreme genomes. Proc Natl Acad Sci U S A. 102:3354–3359.
Martchenko M, Levitin A, Hogues H, Nantel A, Whiteway M. 2007. Transcriptional rewiring of fungal galactose-metabolism circuitry. Curr Biol. 17:1007–1013.
Moses AM, et al. 2006. Large-scale turnover of functional transcription factor binding sites in Drosophila. PLoS Comput Biol. 2:e130.
Nishi K, et al. 1995. The GCR1 requirement for yeast glycolytic gene expression is suppressed by dominant mutations in the *SGC1* gene, which encodes a novel basic-helix-loop-helix protein. Mol Cell Biol. 15:2646–2653.
Odom DT, et al. 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nat Genet. 39:730–732.
Otto W, et al. 2009. Measuring transcription factor-binding site turnover: a maximum likelihood approach using phylogenies. Genome Biol Evol. 2009:85–98.
Perez JC, Groisman EA. 2009a. Evolution of transcriptional regulatory circuits in bacteria. Cell. 138:233–244.
Perez JC, Groisman EA. 2009b. Transcription factor function and promoter architecture govern the evolution of bacterial regulons. Proc Natl Acad Sci U S A. 106:4319–4324.
Piskur J, Rozpedowska E, Polakova S, Merico A, Compagno C. 2006. How did Saccharomyces evolve to become a good brewer? Trends Genet. 22:183–186.
Pollard DA, Moses AM, Iyer VN, Eisen MB. 2006. Detecting the limits of regulatory element conservation and divergence estimation using pairwise and multiple alignments. BMC Bioinformatics. 7:376.
Prabhakar S, et al. 2006. Close sequence comparisons are sufficient to identify human cis-regulatory elements. Genome Res. 16: 855–863.
Schmidt D, et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science. 328: 1036–1040.
Siepel A, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034–1050.
Simpson P, Ayyar S. 2008. Evolution of cis-regulatory sequences in Drosophila. Adv Genet. 61:67–106.
Tanay A, Regev A, Shamir R. 2005. Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. Proc Natl Acad Sci U S A. 102:7203–7208.
Tatusov RL, et al. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 4:41.
Tautz D. 2000. Evolution of transcriptional regulation. Curr Opin Genet Dev. 10:575–579.
Tsong AE, Miller MG, Raisner RM, Johnson AD. 2003. Evolution of a combinatorial transcriptional circuit: a case study in yeasts. Cell. 115:389–399.
Tsong AE, Tuch BB, Li H, Johnson AD. 2006. Evolution of alternative transcriptional circuits with identical logic. Nature. 443:415–420.
Tuch BB, Galgoczy DJ, Hernday AD, Li H, Johnson AD. 2008. The evolution of combinatorial gene regulation in fungi. PLoS Biol. 6:e38.
Tuch BB, Li H, Johnson AD. 2008. Evolution of eukaryotic transcription circuits. Science. 319:1797–1799.
Wang J, Keightley PD, Johnson T. 2006. MCALIGN2: faster, accurate global pairwise alignment of non-coding DNA sequences based on explicit models of indel evolution. BMC Bioinformatics. 7:292.
Ward LD, Bussemaker HJ. 2008. Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. Bioinformatics. 24:i165–i171.
Weirauch MT, Hughes TR. 2010. Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. Trends Genet. 26:66–74.
Woolfe A, et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. 3:e7.
Zheng W, Zhao H, Mancera E, Steinmetz LM, Snyder M. 2010. Genetic analysis of variation in transcription factor binding in yeast. Nature. 464:1187–1191.

Associate editor: George Zhang

Genome Biology and Evolution SMBE